Нечеткое хэширование для детектирования спам-писем

Марченко Алексей
Сергеевич





Agenda

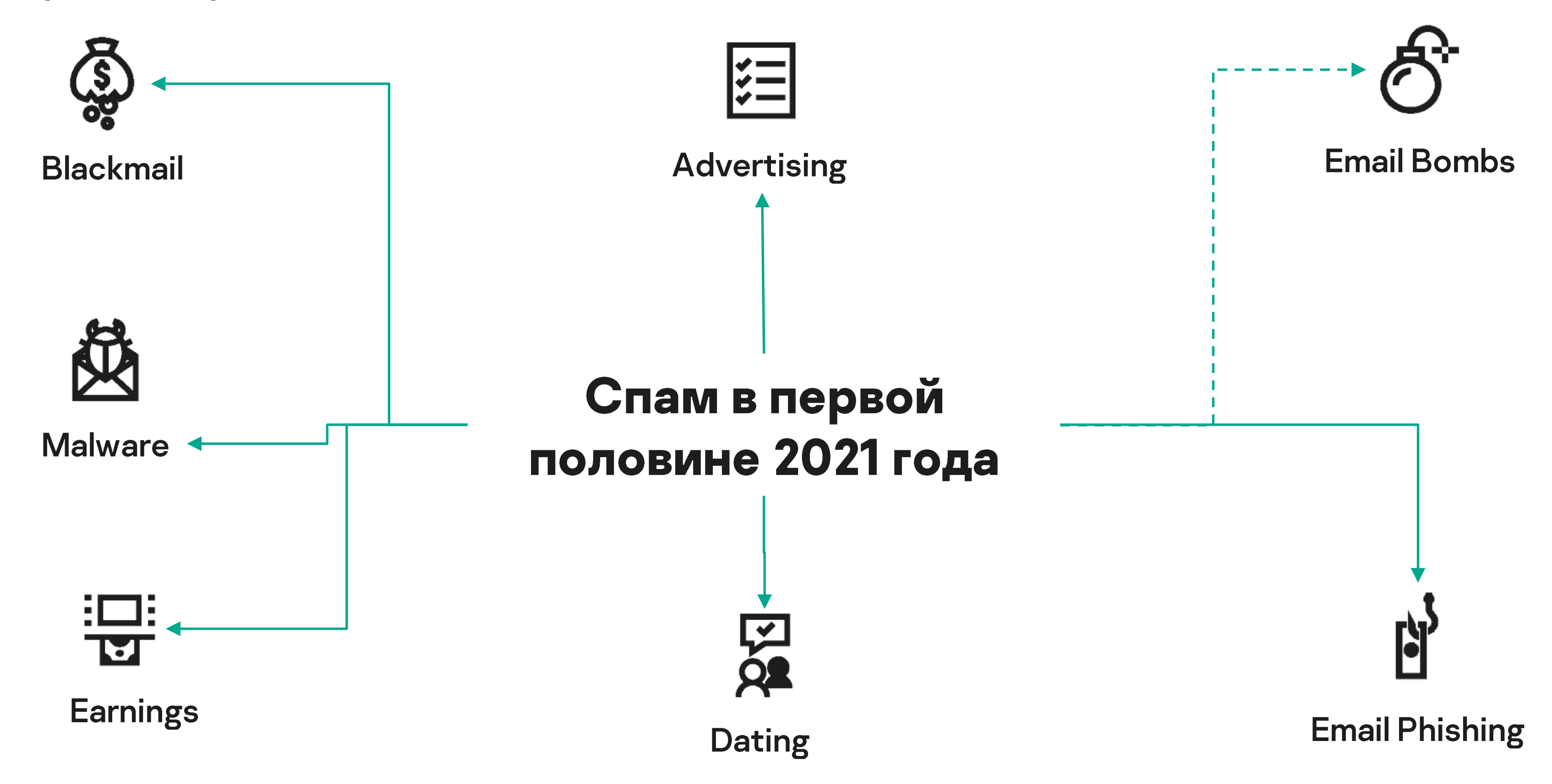
Спам – это проблема?

Если коротко – да.

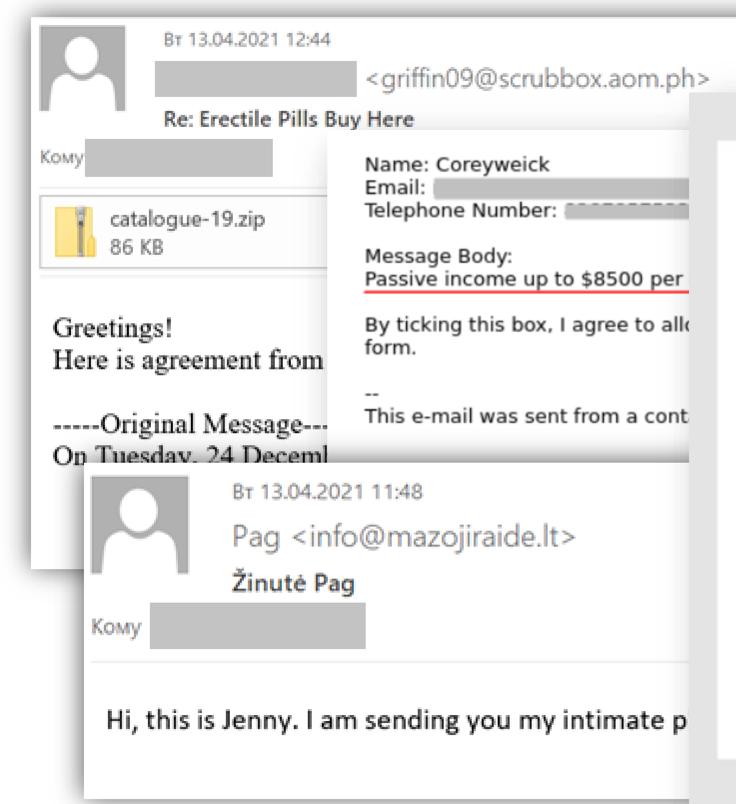
Спам – массовое и не запрошенное

История появления
1978 год. Гэри Тюерк, маркетолог Digital Equipment Corporation рассылает рекламу их нового продукта.

Цели спама
- Реклама
- Мошенничество и фишинг
- Распространение malware
- DDoS
- …


Blackmail
Advertising
Email Bombs
Malware
Спам в первой половине 2021 года
Earnings
Dating
Email Phishing

Примеры спам-писем



Open in your web browser

YOUR PARCEL WILL BE DELIVERED TODAY

Package: 92047387025

Failed delivery attempt: 21/03/2021

Your parcel was returned to our depot and you need to reschedule your package delivery.

To receive your package, we ask that you send us your correct address and pay the new shipping costs "1.99$" at the following link:

COMPLETE MY DELIVERY ADDRESS

Sent by **Notifications**
DPT
If you wish to unsubscribe, please click here.

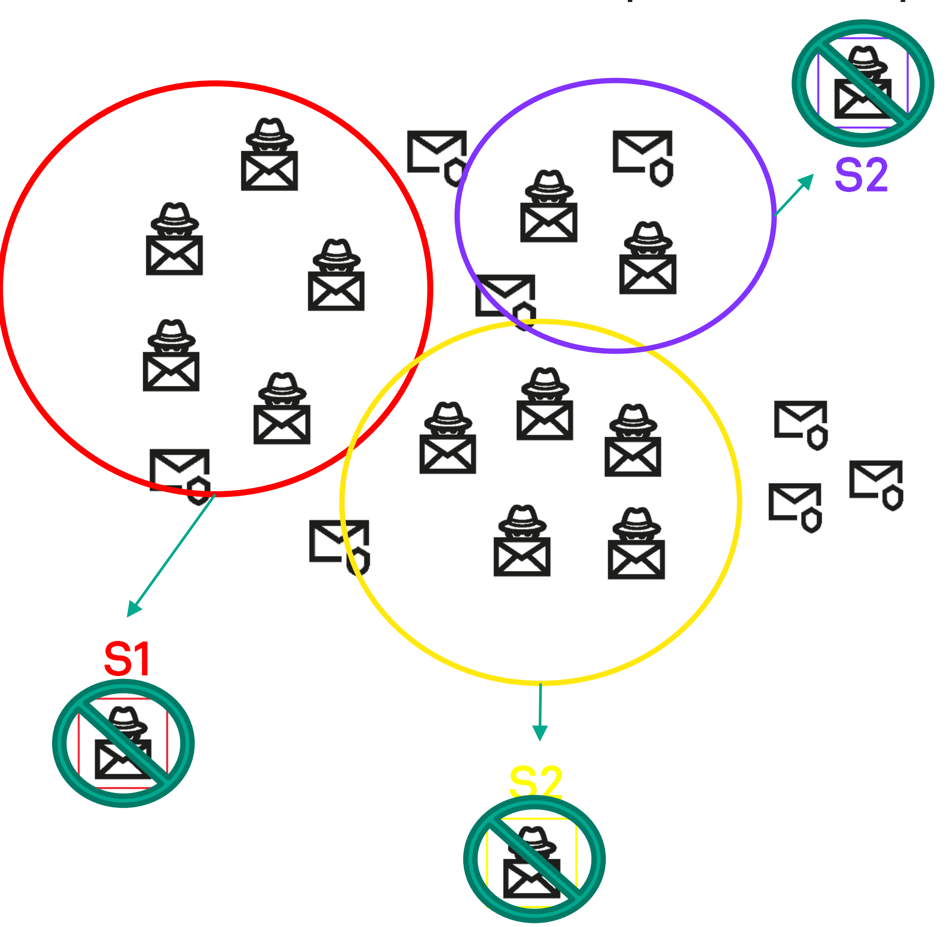
Ущерб от спама

- Нарушает непрерывность бизнеса переполняя почтовые ящики сотрудников. Приводит к повышению времени на операционные задачи.
- Приносит с собой фишинговые ссылки и malware - это приводит к прямым и непрямым финансовым потерям.
- Средний процент спама в почтовом трафике в 2020 году составил 50.7%.
- Согласно публичным отчетам, ущерб, нанесенный спамом коммерческим организациям, исчисляется миллиардами в год.

Методы детектирования спама

Просто берешь… и ловишь.

Общая схема методов детектирования спам-рассылок



1. Сбор данных
2. Выделение рассылок
3. Фильтрация легитимных писем
4. Анализ и выделение характерных признаков каждой спам-рассылки
6. Формирование сигнатур для детектирования аналогичных писем
7. Применение сигнатур

Структура электронного сообщения



IP: 11.22.33.44
MAIL FROM:
<bob@spammer.com>

...
Return Path: <return@nowhere.com>
Message-ID: <ab3sd45ase2@example.com>
Content-Language: en-US
Content-Type: text/plain
X-Mailer: Microsoft Office Outlook, Build 12.0.4210
...

Hello there!

You have won 1 000 000 $ in lottery! To take your money, please, follow this link http://hacked1-domain.com/download-malware123

Информация об отправителе
Сервер, который отослал сообщений (IP адрес, SMTP FROM, …)

Заголовки письма
Техническая информация добавляемая к каждому письму (msgid, date, content type, …).

Тело письма
Контент письма, отображаемый получателю (text, attachments, pictures, …).

Заголовки письма

Subject: I want to steal your personal data!
From: sender@foo.com
To: me@test.com
Date: Mon, 23 Sep 2019 17:00:14 +0300
Message-Id: <h5ced853647da4fd3689a26db412fa4c1@foo.com>
Content-type: multipart/mixed; boundary="===============6411753208318154896=="
X-Mailer: Microsoft Windows Live Mail 14.0.8117.416

Заголовки

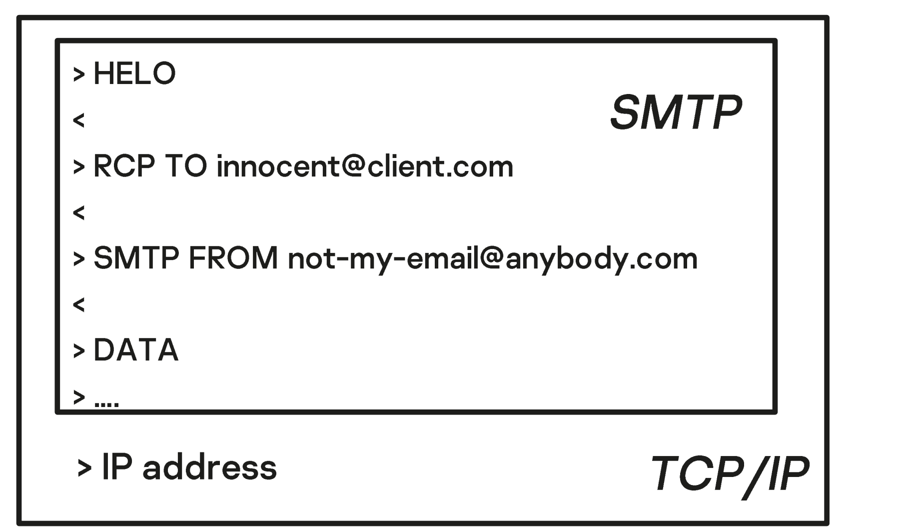
Заголовки описывают как информацию, отображаемую пользователю, так и техническую информацию о письме.

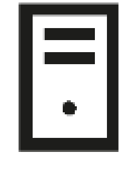
Детектирование спама

Правильно подобранная комбинация заголовков позволяет уникально идентифицировать рассылку.

Информация об отправителе







Информация об отправителе

На уровне TCP/IP нам доступен IP-адрес

На SMTP-уровне доступен SMTP FROM-адрес

Детектирование спама

Позволяет строить списки IP-адресов, рассылающих спам

Технология SPF, позволяет установить подлинность SMTP FROM

Тело письма

Dear recipient!

You have w0n a lottery!
Please come to the link I send in the attachment to get your 10 000 $ PRIZE! !!!

Best regards,
Some Very Popular Governmental Lottery

Тело письма

Содержит текст, который можно проанализировать

Содержит вложения и картинки

Детектирование спама

Текст является мощным признаком для определения спам-рассылок

Почему анализа заголовков и отправителя недостаточно

Спам через формы обратной связи
Спам-текст вставляется в поле «комментарий», а адрес жертвы в поле контактного адреса.

Backscatter
Используется легитимный, плохо настроенный сервер и механизм NDR для отправки сообщений от имени сервера

Web Mail / Cloud
Спам отправляется с сервера cloud-платформы

Почему текст спам-писем тяжело анализировать «как текст»

Зашумления
Намеренные ошибки, typesquatting, избыточная пунктуация и т.д.

Перебор синонимов
В разных сообщениях одной рассылки используются слова-синонимы или «условно» подходящие по контексту

Языки
Спам отправляется на разных языках и потребуется целый штат переводчиков

Нечеткое хэширование и кластеризация

Очень четко.

На этом примере можно показать, как работает нечеткое хэширование, и на этом же примере будет понятно, почему для него выполняется правило «чем сильнее отличаются исходные данные, тем сильнее отличаются хэши»

CTPH = ab4e + efb4 + b4c3 = ab4eefb4b4c3

RollingHashValue DIV 10 = 0 → «разрезаем»

На этом примере можно показать, как работает нечеткое хэширование, и на этом же примере *несмотря на отличия в тексте* будет понятно, почему для него выполняется правило «чем сильнее отличаются исходные данные, тем сильнее отличаются хэши»

CTPH* = ab4e + *fc5e* + b4c3 = ab4efc5eb4c3

Fuzzy Hashing
«Чем сильнее отличаются исходные данные, тем сильнее отличаются хэши»

Реализация CTPH
Rolling Hash для определения точек «разреза» + Traditional Hash для фиксации значения каждого кусочка информации

Дистанция Левенштейна

$$\operatorname{lev}(a,b)=\begin{cases}|a| & \text{if } |b|=0,\\ |b| & \text{if } |a|=0,\\ \operatorname{lev}(\operatorname{tail}(a),\operatorname{tail}(b)) & \text{if } a[0]=b[0]\\ 1+\min\begin{cases}\operatorname{lev}(\operatorname{tail}(a),b)\\ \operatorname{lev}(a,\operatorname{tail}(b))\\ \operatorname{lev}(\operatorname{tail}(a),\operatorname{tail}(b))\end{cases} & \text{otherwise.}\end{cases}$$

Дистанция Левенштейна как функция схожести СТРН-хэшей

Дистанция Левенштейна является метрикой (с математической точки зрения)

Применение fuzzy-хэширования



Системы «Антиплагиат»



Детектирование malware



Поиск похожих изображений

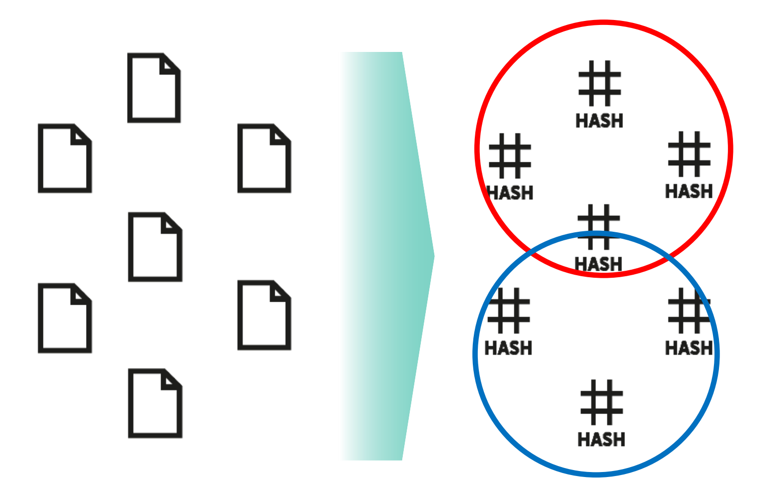
Предпосылки идеи

Thank you so much for choosing our online streaming solutions We are always happy to see you being part of our warm and friendly team! it is time to move to premium subscription! You'll be automatically shifted to the premium as soon as the trial period expires. Thank you so much for giving your settlement details. Below are your order info for order number:M0082261491993908 BravoMovies Premium subscription Four weeks: $39.99* *All fees are integrated in the price total In case you have any thoughts about your current subscription or perhaps would want to cancel it, give us a call at: +1(213)-267-7082 You are actually helping us donate a part of our sales to the COVID-19 Reaction Organization. We think it is crucial to support other individuals! All the best BravoMovies.

Thanks for choosing our internet streaming service We are extreamely happy to see you becoming part of our friendly crew! it is now time to move to premium subscription You will be digitally moved to the premium as soon as the free period ends. Thank you for providing your payment details. Below are your order info for transaction id:M0082213060303748 BravoMovies Premium Four weeks: $39.99* *All fees are integrated in the price balance If perhaps have some concerns about your premium or perhaps want to cancel it, contact us at: +1(213)-267-7082 You are helping us deliver portion of our sales to the COVID-19 Reaction Organization. We really think it is important to support other individuals! Always yours, BravoMovies.

Thanks a lot for selecting our internet streaming solutions! We are extreamely happy to witness you becoming part of our warm and friendly crew! it is now chance to proceed to premium You will be digitally shifted to the premium as soon as the trial stage expires Thank you so much for giving your monthly payment info Listed here are your order information for transaction no:M0082282634283918 BravoMovies Premium subscription Four weeks: $39.99* *All fees are integrated in the amount balance If perhaps have any concerns about your current membership or would want to cancel it, call us at: +1(213)-267-7082 You are actually helping us give away a part of our sales to the COVID Solution Organization. We think it's crucial to support others! Best regards BravoMovies.

Идея



Использовать нечеткое хэширование, чтобы «сгладить» различия в данных

Кластеризовать fuzzy-хэши, чтобы выделить спам-рассылки и отсеять легитимные письма

Алгоритм не должен опираться на точки за пределами исходных данных

Присутствует шум (легитимные письма)

DBSCAN

Неизвестно количество кластеров

Кластеры не имеют четкой «формы»

DBSCAN

Плотностной алгоритм кластеризации, основанный на выделении компонент связанности

Позволяет находить кластеры произвольной «формы»

Не требует считать промежуточные точки в пространстве

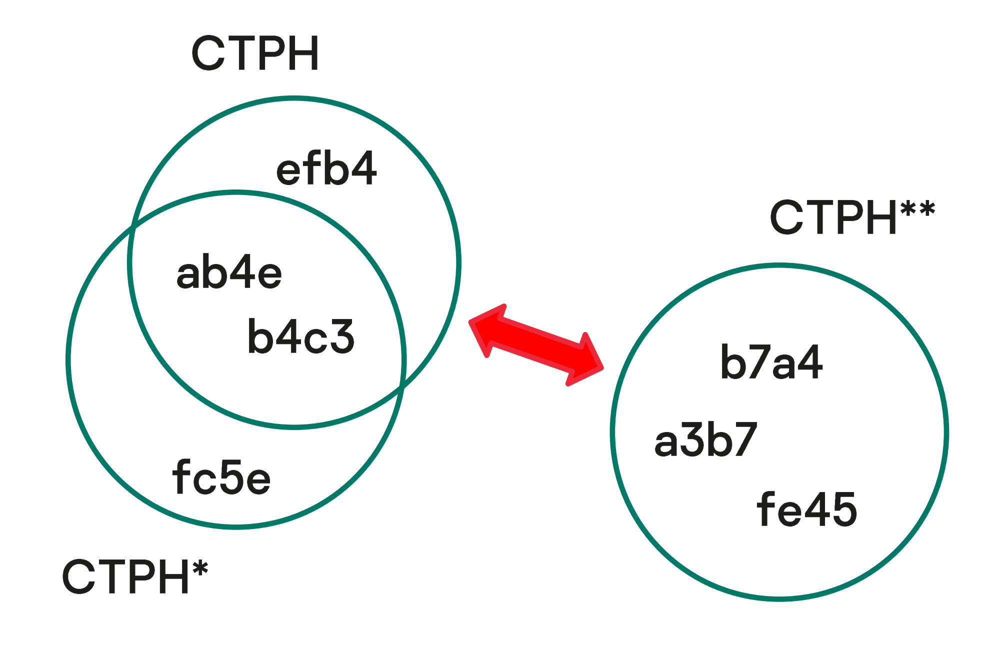
Прекрасно умеет работать с зашумленными данными

Оптимизация количества вычислений

CTPH = ab4e + efb4 + b4c3 = ab4eefb4b4c3

CTPH* = ab4e + fc5e + b4c3 = ab4efc5eb4c3

CTPH** = b7a4 + a3b7 + fe45 = b4c4a3b7fe45



У двух хэшей нет ни одного общего «кусочка» => у исходных данных нет ничего общего

Позволяет разделить исходные данные на непересекающиеся группы и кластеризовать отдельно => меньше подсчетов расстояний

Эксперимент

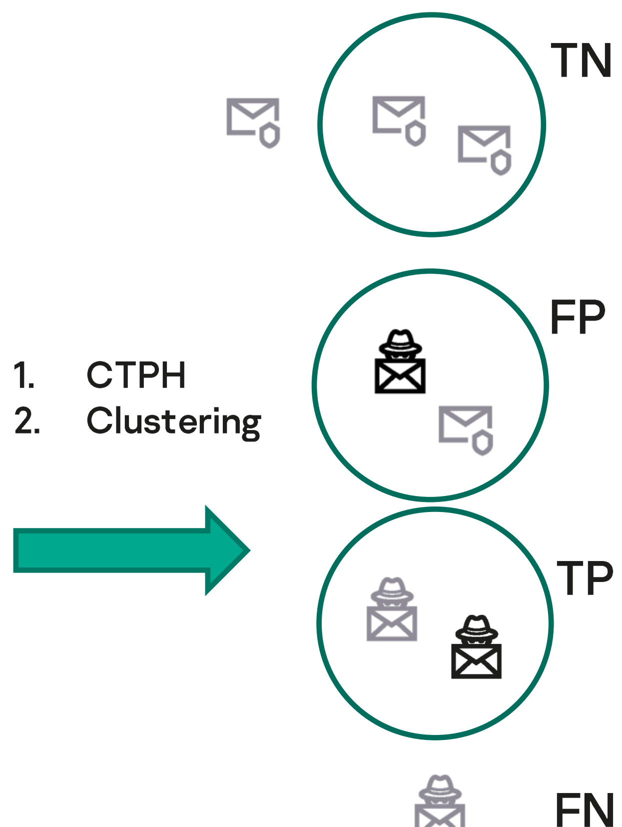
100 000 помечаем как известный спам

45 000 помечаем как неизвестные

45 000 помечаем как неизвестные

145 000 спам-писем

45 000 легитимных писем



Результаты эксперимента

DBSCAN *Eps*	Detection Precision	Detection Recall	FP count
X	1	0.2033	0
X+10	1	0.3054	0
X+20	1	0.4981	0
X+30	0.9998	0.6320	2
X+40	0.9996	0.74025	5
X+50	0.9992	0.78075	11

Расходимся?

Архитектура

Еще раз… почему не кластеризовать письма прямо у клиента?

Общая архитектура защитных решений



Считать хэши и кластеризовать в клиентской среде нельзя

Нет необходимых данных

Нет времени

Нет необходимых мощностей

Неудобно контролировать качество

Идея спам-терминов

Thank you so much for choosing our online streaming solutions! We are always happy to see you being part of our warm and friendly team! it is time to move to premium subscription! You'll be automatically shifted to the premium as soon as the trial period expires. Thank you so much for giving your settlement details. Below are your order info for order number:M0082261491993908 BravoMovies Premium subscription Four weeks: $39.99* *All fees are integrated in the price total In case you have any thoughts about your current subscription or perhaps would want to cancel it, give us a call at: +1(213)-267-7082 You are actually helping us donate a part of our sales to

Thanks for choosing our internet streaming service! We are extreamely happy to see you becoming part of our friendly crew! it is now time to move to premium subscription! You will be digitally moved to the premium as soon as the free period ends. Thank you for providing your payment details. Below are your order info for transaction id:M0082213060303748 BravoMovies Premium Four weeks: $39.99* *All fees are integrated in the price balance If perhaps have some concerns about your premium or perhaps want to cancel it, contact us at: +1(213)-267-7082 You are helping us deliver portion of our sales to the COVID-19 Reaction Organization. We really think it is important t

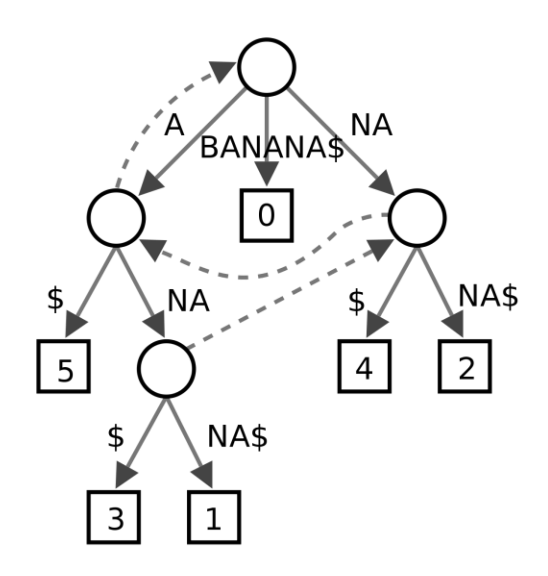
Thanks a lot for selecting our internet streaming solutions! We are extreamely happy to witness you becoming part of our warm and friendly crew! it is now chance to proceed to premium! You will be digitally shifted to the premium as soon as the trial stage expires. Thank you so much for giving your monthly payment info. Listed here are your order information for transaction no:M0082282634283918 BravoMovies Premium subscription Four weeks: $39.99* *All fees are integrated in the amount balance If perhaps have any concerns about your current membership or would want to cancel it, call us at: +1(213)-267-7082 You are actually helping us give away a part of our sales to the COVID Solution Organization. We think it's crucial to support others! Best regards, BravoMovies.



Four weeks: $39.99* *All tax are integrated in the

- Легковесно
- Легко применять
- Легко интерпретировать

LCS



Определение
LCS (Longest Common Substring), семейство алгоритмов поиска максимальной общей подстроки в заданном множестве текстов

Реализация LCS
Существует множество реализаций, одним из наиболее быстрых является вариант на генерализованных суффиксных деревьях

Хорошие и плохие подстроки

Four weeks: $39.99* *All tax are integrated in the

Please let me know if you have any questions Thanks

Не каждая максимальная подстрока характеризует спам-рассылку

Подстроки, характеризующие спам-рассылку, будем называть «спам-термины»

Вопрос: как определить, является ли подстрока спам-термином?

Фильтрация подстрок

По виду подстроки

- Малая длина
- Малое количество слов
- Отсутствие идентификаторов

По коллекциям

- Присутствие в других кластерах (спам-рассылках)
- Присутствие в коллекциях легитимных писем

DBSCAN *Eps*	Detection Precision	Detection Recall	FP count	%, valid spam terms
X	1	0.2033	0	65.87
X+10	1	0.3054	0	72.10
X+20	1	0.4981	0	56.66
X+30	0.9998	0.6320	2	47.38
X+40	0.9996	0.74025	5	45.54
X+50	0.9992	0.78075	11	45.21

Архитектура алгоритма

Конвертация в CTPH → **Кластеризация** → **Извлечение LCS** → **Фильтрация**

Конвертация в CTPH

- Сглаживает зашумление
- «Цепляется» за повторяющиеся фрагменты
- Убирает необходимость анализировать тексты

Кластеризация

- Выделяет спам-рассылки
- Отфильтровывает легитимные письма

Извлечение LCS

- Формирует простую сигнатуру

Фильтрация

- Позволяет отфильтровать «плохие» сигнатуры и получить спам-термины

Глобальная архитектура



Сильные и слабые стороны

Сильные стороны

- Дешевизна трафика
- Простота интерпретации
- Скорость детектирования в клиентской среде

Слабые стороны

- Временные затраты на получение спам-терминов
- Большие вычислительные ресурсы в инфраструктуре

Результаты и дальнейшие планы

И что, взлетело?

01.10.2020



Запуск технологии

58 000+



Спам-терминов выпущено

25 000 000+



Спам-писем заблокировано и у пользователей

Пример

Greeting's Beloved I am Mrs Lizzy Febian I'm 63 Years old. I am a widow, sufferings from a lon g time cancer disease. I want to use my Inherited Fund to help the orphans and Widows. If you are willingly to carry out this work for God's sake Contact me for more details Thanks and God bless you. your Beloved sister Mrs Lizzy Febian

--

I am Mrs. Monika Da

I am a widow, sufferings from a long time cancer disease.

I want use my Inherited Fund to Donate for the Orphans Needy and Widows.

If you are willingly to carry out this work for God's sake.

Contact me for more details.

With best regards,
E-mail: monika.1274
Mrs Monika David.

Pozdrowienia od pani Lizzy Febian Nazywam się Pani Lizzy Febian Mam 63 lata. Jestem wdową i od dawna cierpię na raka. Chcę wykorzystać swój odziedziczony fundusz, aby pomóc sierotom i wdowo m. Jeśli zechcesz wykonać tę pracę na miłość boską, skontaktuj się ze mną, aby uzyskać więcej informacji Dziękuję i niech cię Bóg błogosławi. twoja ukochana siostra Pani Lizzy Febian Greet ing's From Mrs Lizzy Febian I am Mrs Lizzy Febian I'm 63 Years old. I am a widow, sufferings from a long time cancer disease. I want to use my Inherited Fund to help the orphans and Widow s. If you are willingly to carry out this work for God's sake Contact me for more details Than ks and God bless you. your Beloved sister Mrs Lizzy Febian

Пример

Здравствуйте, позвольте мы ознакомим Вас с нашей програм
UltimateSpiderBot - Программа для быстрого продвижения
Миллионы уникальных посещений Вашего сайта.
Результат:
? Ваш сайт в ТОП поисковых систем. ? Счетчик посещений
о всем показателям. ? Зарабатывайте на рекламе.
У программы существует возможность скликивать сайты конк
ыдаче.
Простыми словами, программа выведет Ваши сайты в ТОП, а
ии.
Возможно Ваши конку

Вам интересно вывест

{http://bit.ly/free-
-fast}-
Пожизненная лицензи

Здравствуйте, так как Вы владеете в
UltimateSpiderBot - Программа для
мы.
МИЛЛИОНЫ УНИКАЛЬНЫХ ПОСЕЩЕНИЙ ВАШИХ
Пожизненная лицензия!

У программы существует возможность
ыдаче.
Простыми словами, программа выведет Ваши сайты в ТОП засчитанные дни, а сайты конкурентов пон
изит ниже плинтуса…

Возможно Ваши конкуренты уже используют наш софт и выводят свои проекты в топ…

Вам интересно вывести свой сайт в ТОП, без вложений, за несколько дней?

{http://bit.ly/freetopfast}- Подробнее о программе UltimateSpiderBot {http://bit.ly/freetopfas
t}-

Здравствуйте, позвольте мы ознакомим Вас с нашей программой.
UltimateSpiderBot - Программа для быстрого продвижения Веб-сайтов.
Миллионы уникальных посещений Вашего сайта.

Результат:
? Ваш сайт в ТОП поисковых систем.
? Счетчик посещений растет на глазах.
? Высокие рейтинги по всем показателям.
? Зарабатывайте на рекламе.

У программы существует возможность скликивать сайты конкурентов для понижения их в поисковой в
ыдаче.
Простыми словами, программа выведет Ваши сайты в ТОП, а сайты конкурентов потеряют свои позици
и.
Возможно Ваши конкуренты уже используют наш софт и выводят свои проекты в топ…

Подробнее на нашем Веб-сайте.
https://freetopfast.com/

To view all of this form's submissions, visit https://www.panasonicaircon.co.nz/index.php/dash
board/reports/forms?qsid=1522012625

Пример

美国纯电池专线 材积除6000（香港海运+美国UPS/FEDEX派送服务）含燃油含税

分区 71kg+ 101kg+ 501kg+ 10

美国全境48州 (0-9开头邮编) 19

以上是全包价：接受纯电池/移动电源/电

王锋 Jayleke Manager

深圳市飞腾运通货运代理有限公司

ADD:深圳市南山区创业路亿利达大厦A座

Tel:0755-27872162 Fax:0755-278
企业QQ: 1941553703

Mob:189 2745 4613（微信同号）

http:{http://www.szftyt.com/}w

深圳市飞腾运通货运代理有限公司

美国专线-双清包税

邮编分区 21KG+ 75KG+ 201KG+ 3
viewfile?f=F55E9CA39EC25F116E2ED7BF0
A4886E1FAA9E265B465AC24F8B0C2D072DB3
93A2F52593564C6&mailid=ZL2811-iK~I1B
积除6000 5-7工作日提取

美国全境48州 (0-9开头邮编) * *

可接一般贸易报关

王锋 Jayleke Manager

深圳市飞腾运通货运代理有限公司

ADD:深圳市南山区创业路亿利达大厦A座8A19

Tel:0755-27872162 Fax:0755-27872161

Mob:189 2745 4613（微信同号） 企业QQ: 1941553703
http:{http://www.szftyt.com/}www.szftyt.com

美国空派大促销

{cid:_Foxmail.1@4aac9f5f-f970-4de4-b7b1-fb9a4e89b945}美国专线-促销(大陆飞)

邮编分区 21KG+ 75KG+ 201KG+ 301KG+ 500KG+ 非固定上网地点{about:Attach/128280(05-07-12-00-52).png} 5-7工作日提取

美国全境48州 (0-9开头邮编) * * 54 53 52 洛杉矶 芝加哥 纽约市

企业QQ: 1941553703
微信: 18927454613

王锋 Jayleke Manager

深圳市飞腾运通货运代理有限公司

ADD:深圳市南山区创业路亿利达大厦A座8A19

{<REDACTED_LINK>}Tel:0755-27872162 Fax:0755-27872161

Mob:189 2745 4613

http:www.szftyt.com

Что дальше?



Весовые спам-термины
В ситуациях, когда есть большое количество «плохих» общих подстрок, которые в совокупности дают «хорошую» сигнатуру



Не самые длинные подстроки
В ситуациях, когда самая длинная подстрока оказалась «плохой» сигнатурой, но следующая за ней по длине – валидный спам-термин

Спасибо!

... и
время вопросов